Decoupled Spatio-Temporal Consistency Learning for Self-Supervised Tracking

Yaozong Zheng[1,2], **Bineng Zhong**[1,2*], **Qihua Liang**[1,2], **Ning Li**[1,2], **Shuxiang Song**[1,2]

[1]Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China

[2]Guangxi Key Lab of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, China

yaozongzheng@stu.gxnu.edu.cn, bnzhong@gxnu.edu.cn, qhliang@gxnu.edu.cn

ningli65536@mailbox.gxnu.edu.cn, songshuxiang@mailbox.gxnu.edu.cn

Abstract

The success of visual tracking has been largely driven by datasets with manual box annotations. However, these box annotations require tremendous human effort, limiting the scale and diversity of existing tracking datasets. In this work, we present a novel Self-Supervised Tracking framework named **SSTrack**, designed to eliminate the need of box annotations. Specifically, a decoupled spatio-temporal consistency training framework is proposed to learn rich target information across timestamps through global spatial localization and local temporal association. This allows for the simulation of appearance and motion variations of instances in real-world scenarios. Furthermore, an instance contrastive loss is designed to learn instance-level correspondences from a multi-view perspective, offering robust instance supervision without additional labels. This new design paradigm enables SSTrack to effectively learn generic tracking representations in a self-supervised manner, while reducing reliance on extensive box annotations. Extensive experiments on nine benchmark datasets demonstrate that SSTrack surpasses *SOTA* self-supervised tracking methods, achieving an improvement of more than 25.3%, 20.4%, and 14.8% in AUC (AO) score on the GOT10K, LaSOT, TrackingNet datasets, respectively.

Code — https://github.com/GXNU-ZhongLab/SSTrack

Introduction

Given an arbitrarily initial target, visual object tracking (VOT) requires recognizing and tracking an object in subsequent video frames. To accomplish this computer vision task, current high-performance visual tracking algorithms are typically trained using the full bounding box annotations of published tracking datasets (Fan et al. 2019; Müller et al. 2018; Lin et al. 2014; Huang, Zhao, and Huang 2021), as shown in Fig.1(a). However, the bounding boxes in existing VOT benchmarks rely on tremendous human efforts, making it difficult to expand their scale and diversity, such as the number of arbitrary tracked objects and open tracking scenarios. This poses a challenge for transformer-based tracking algorithms (Ye et al. 2022; Cui et al. 2022; Xing et al. 2023; Zheng et al. 2024), as they tend to be particularly data-hungry. From this perspective, equipping a model with the ability to automatically learn to track instances from unlabeled videos becomes crucial in the field of visual tracking. Therefore, as shown in Fig.1(b), we reconsider the need for box annotations by exploring *a new self-supervised tracking algorithm under an initial bounding box setting*.

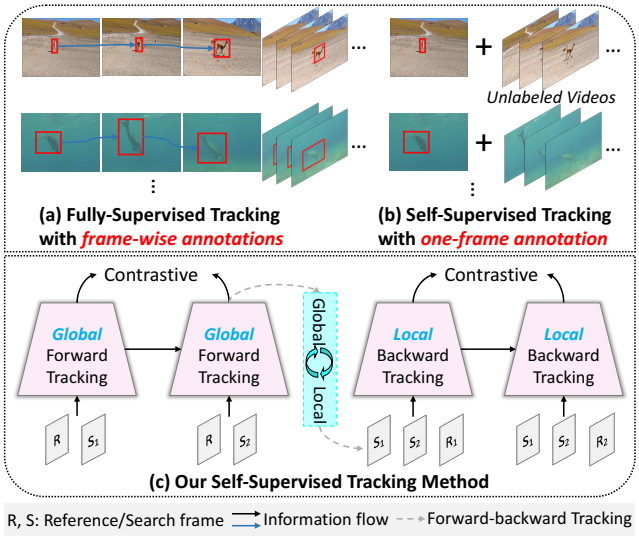

Figure 1: The annotation requirements of different tracking tasks and our proposed framework. (a) The fully-supervised tracking methods (Li et al. 2019; Chen et al. 2021) with frame-wise annotations. (b) The self-supervised tracking methods (Sio et al. 2020; Yuan et al. 2020) with one-frame annotation. (c) Our self-supervised tracking method based on decoupled spatio-temporal consistency training framework and instance contrastive loss.

To minimize the reliance on box annotations, some self-supervised tracking methods (Wang, Jabri, and Efros 2019; Sio et al. 2020; Yuan et al. 2020; Li et al. 2023) have been proposed to learn object correspondences from unlabeled videos. They learn instance tracking representations through contrastive learning or cycle-consistency matching strategies. For instance, S^2SiamFC (Sio et al. 2020) and TADS (Li et al. 2023), based on supervised tracking methods, generate training pairs through random region selection and data augmentation to train self-supervised trackers. Mean-

*Corresponding author.

while, self-SDCT (Yuan et al. 2020) and CycleSiam (Yuan, Wang, and Chen 2020) rely on the principle of cyclic consistency to construct a self-supervised tracking framework with forward-backward alignment. Despite previous studies performing well in most tracking scenarios, they still face a significant performance bottleneck *due to the difficulty in effectively leveraging the rich spatio-temporal context and instance correspondence in continuous video frames*.

In this work, we propose a novel self-supervised visual tracking framework, called **SSTrack**, which aims to eliminate the need for expensive manual annotations while efficiently injecting spatio-temporal contextual information. As shown in Fig.1(c), we reconsider the design of the self-supervised tracking framework from a new perspective. Unlike fully supervised methods (Cai, Liu, and Wang 2024; Zheng et al. 2024; Bai et al. 2024) that capture context through multi-frame inputs, directly learning temporal context in self-supervised tracking is a significant challenge due to the lack of annotated video data. To address this challenge, a decoupled spatio-temporal consistency training framework is introduced to automatically learn rich target information across timestamps. Specifically, we first perform forward tracking, globally searching for the spatial position of object. Then, we conduct backward tracking, locally perceiving the appearance and motion (temporal) changes of the instance. Through this decoupled learning approach, we achieve global spatial localization and local temporal association within a unified framework, thereby effectively utilizing both labeled and unlabeled video data. Furthermore, we introduce an instance contrastive loss function to learn instance-level correspondence across views, providing robust instance supervision without any labels. This new design paradigm enables SSTrack to effectively learn generic tracking representations in a self-supervised manner, while reducing reliance on extensive box annotations. Extensive experiments show that our approach achieves excellent tracking performance with limited annotations and significantly narrows the performance gap between self- and fully supervised tracking methods. The main contributions of this work are as follows.

- We propose a novel self-supervised tracking pipeline named SSTrack, based on a decoupled spatio-temporal consistency training framework. It end-to-end learns cross-frame target representations via global spatial localization and local temporal association.
- We introduce an instance contrastive loss function to learn instance-level correspondence from a multi-view perspective, offering robust instance supervision without any labels.
- Our tracker achieves a new *SOTA* tracking results on nine visual tracking benchmarks, including GOT10K, LaSOT, TrackingNet, $\text{LaSOT}_{\text{ext}}$, VOT2020, TNL2K, VOT2018, UAV123 and OTB100.

Related Work

Fully-Supervised Tracking Methods

The prevailing visual tracking algorithms (Bertinetto et al. 2016; Chen et al. 2020; Yan et al. 2021; Ye et al. 2022) predominantly adhere to the supervised tracking paradigm and achieve high performance by training on large-scale labeled datasets such as LaSOT (Fan et al. 2019), TrackingNet (Müller et al. 2018), COCO (Lin et al. 2014), and GOT10K (Huang, Zhao, and Huang 2021). These supervised tracking algorithms can be broadly categorized into two types: Siamese tracking framework (Bertinetto et al. 2016; Li et al. 2018; Chen et al. 2020; Cheng et al. 2021; Zhang and Peng 2019) and Transformer tracking framework (Ye et al. 2022; Cui et al. 2022; Chen et al. 2022; Xing et al. 2023; Zheng et al. 2024). The former typically follows a three-stage approach involving feature extraction, fusion, and bounding box prediction for visual tracking, while the latter generally employs a transformer network to simultaneously perform feature extraction and fusion.

Benefiting from training datasets with thousands of manual bounding box annotations, these methods have achieved significant performance gains. However, constructing large-scale video datasets is exceedingly time-consuming and costly, making it challenging to keep pace with rapid advancements in supervised tracking algorithms and often leading to gaps between data distributions in real-world scenarios. Essentially, the availability of large-scale, high-quality datasets is increasingly becoming a bottleneck for the progress of supervised tracking. Thus, there is an urgent need to research a novel and effective self-supervised tracking framework to alleviate this problem.

Self-Supervised Tracking Methods

Unlike mainstream fully-supervised tracking algorithms, self-supervised tracking algorithms (Sio et al. 2020; Yuan et al. 2020; Li et al. 2023) face greater challenges due to the lack of sufficient supervision signals. Current self-supervised tracking frameworks are typically divided into those based on cycle consistency and those based on contrastive learning methods. 1) *Cyclic-consistency based self-supervised tracking methods*. self-SDCT (Yuan et al. 2020) introduces a multi-cycle consistency loss and low similarity dropout strategy to train the feature extraction network, enhancing the robustness of the self-supervised tracker. CycleSiam (Yuan, Wang, and Chen 2020) leverages cycle-consistent techniques, along with region proposal and mask regression networks, to explore a Siamese self-supervised tracking framework that simultaneously performs tracking and segmentation tasks. 2) *Contrastive learning based self-supervised tracking methods*. S^2SiamFC (Sio et al. 2020) randomly selects a region of the image and a corresponding enlarged region as a sampling pair, and proposes adversarial appearance masking technique for self-supervised tracking. However, this training strategy tends to sample low-quality sample pairs and fails to utilize temporal information from multiple consecutive frames. TADS (Li et al. 2023) proposes a generalized data augmentation technique such as crop-transform-paste operation and is based on several supervised tracking frameworks (Li et al. 2019; Chen et al. 2021) to train high performance self-supervised trackers.

Inspired by these studies, researching visual tracking algorithms with minimal supervision signals emerges as a highly promising direction. However, unlike these works,

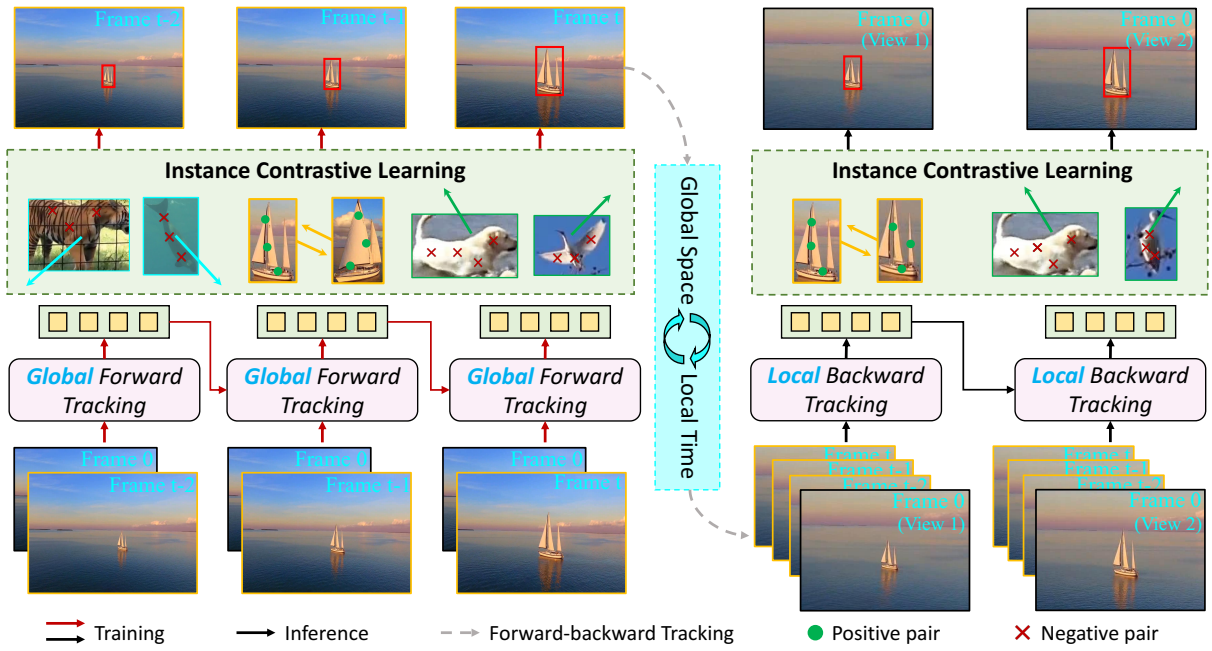

Figure 2: SSTrack training and inference pipeline. 1) Forward Tracking: Given an initial frame and a global search frame, our method performs a global search to identify potential target locations. 2) Backward Tracking: We apply local cropping and data augmentation to the original image pairs, generating two video clips with different views as inputs to our model. This simulates the diverse appearance changes of target in real-world scenarios. 3) Instance Contrastive Learning: An instance contrastive loss is introduced to learn the similarity between different instances, achieving a robust instance tracking representation.

we introduce a new self-supervised training framework from the perspective of decoupled spatio-temporal modeling, which avoids the impact of low-quality sample pairs on the training process. Furthermore, we propose a novel baseline method named SSTrack, focusing on unlocking the potential of self-supervised tracking by collecting and correlating target information across timestamps and views.

Methodology

In this section, we first revisit the definition of the self-supervised visual tracking task and briefly introduce our SSTrack framework. We then provide a detailed description of SSTrack's two components: the decoupled spatio-temporal consistency training framework and the instance contrastive loss.

Self-Supervised Pipeline: SSTrack

Task Definition. Given the initial information (bounding box annotation) of any instance in the first frame, the self-supervised visual tracking task aims to *train a tracker from completely unlabeled videos* and accurately locate the target in subsequent video frames.

Framework Formulation. In order to comprehensively understand our novel self-supervised tracking framework, it is necessary to summarize the previous mainstream fully supervised tracking methods (Li et al. 2019; Ye et al. 2022; Xing et al. 2023; Zheng et al. 2024).

Despite differences in technical solutions, nearly all top-performing fully supervised methods are based on a common principle: embedding paired frame and bounding box, such as (I_r, B_r), into the tracking network $\mathcal{E}$. As a result, we summarize the fully supervised tracking as follows:

$$B_s = \mathcal{E}\left(I_s, \{(I_r, B_r)\}_n\right), \tag{1}$$

where B_s is the bounding box predicted for a given search frame I_s. $\{(I_r, B_r)\}_n$ represents a pair (initial) or multiple pairs of frames and bounding boxes. Then, $\{(I_r, B_r)\}_n$ are used to guide the localization of the search frame I_s.

Although the field of visual tracking is dominated by fully-supervised algorithms, exploring visual tracking algorithms based on other or minimal supervision signals is a highly promising research direction, as it offers the potential to eliminate dependency on labeled data. Therefore, as shown in Fig.2, we introduce SSTrack, a new self-supervised tracking method based on a decoupled spatio-temporal consistency training framework. Meanwhile, we propose a simple and effective instance contrastive loss to achieve high-performance self-supervised tracking.

Theoretically, our self-supervised framework jointly learns bounding box decoding and general tracking representation from unlabeled videos. It not only leverages the powerful idea of cyclic consistency training strategy, but also inherits the advantages of contrastive learning in general representation learning. Specifically, our self-supervised solution is divided into two stages: forward tracking and backward tracking. To easily understand the self-supervised

tracking process, we use $\mathcal{V}$ to denote the specific target context learned from the reference $\{(I_r, B_r)\}_n$, and then we iteratively employ tracking network $\mathcal{E}$ i times in a forward manner from time step $t-i$ to t:

$$\begin{aligned} B_s^t &= \mathcal{E}^i(I_s^t, \mathcal{V}) \\ &= \mathcal{E}(I_s^{t-1}, \mathcal{E}(I_s^{t-2}, ..., \mathcal{E}(I_s^{t-i}, \mathcal{V}))). \end{aligned} \tag{2}$$

In the backward tracking phase, the tracker $\mathcal{E}$ is employed backwards i times from time step t to $t-i$:

$$\begin{aligned} B_s^{t-i} &= \mathcal{E}^i(I_s^{t-i}, \mathcal{V}) \\ &= \mathcal{E}(I_s^{t-i+1}, \mathcal{E}(I_s^{t-i+2}, ..., \mathcal{E}(I_s^t, \mathcal{V}))). \end{aligned} \tag{3}$$

Based on the above formulation, we can construct a self-supervised tracking framework. We further develop optimization objectives for the proposed model to effectively learn the target correspondence from unlabeled videos. Specifically, we choose classification and regression losses as our optimization objectives. The tracking optimization objective $\mathcal{L}_{track}$ can be formulated as:

$$\mathcal{L}_{track} = \mathcal{L}_{cls}(B_s, B_s^{gt}) + \mathcal{L}_{reg}(B_s, B_s^{gt}), \tag{4}$$

where $\mathcal{L}_{cls}$ is the Focal loss (Lin et al. 2017) and $\mathcal{L}_{reg}$ denotes the combination of GIoU loss (Rezatofighi et al. 2019) and $\mathcal{L}_1$ loss.

Decoupled Spatio-Temporal Consistency Training Framework

Accurately obtaining target identity information is crucial for a self-supervised tracking framework. A straightforward approach is to randomly crop diverse local regions as training sample pairs, aligning with the configuration of traditional local trackers (Sio et al. 2020). However, this method tends to produce low-quality sample pairs and fails to effectively leverage the rich spatio-temporal context, thus becoming a performance bottleneck for self-supervised tracking algorithms. In contrast, we believe a good self-supervised tracking algorithm should automatically locate target instance from a global region, without being limited to a local tracking setup. As a result, we propose a novel decoupled spatio-temporal consistency training framework that seamlessly switches between global and local tracking, automatically identifying and locating instance.

To simplify the model design, we do not add an additional global tracker but instead use a shared ViT (Dosovitskiy et al. 2021) as the fundamental tracking network. Specifically, we decouple forward-backward tracking into global and local tracking. In the forward tracking stage, given an initial frame $I_r \in \mathbb{R}^{3\times H_r \times W_r}$ and an uncropped/full search frame $I_s \in \mathbb{R}^{3\times H_s \times W_s}$, the ViT receives them and performs joint feature extraction and fusion to globally search for the potential target's spatial location. Based on the current tracking results, we then crop the full search frame online to match the size of the template frame, which serves as the template frame for backward tracking. Simultaneously, we apply data augmentation operations, such as scaling, shearing, and blurring, to the initial frame to generate multiple video frames with different views, which serve as the search frames for backward tracking. This approach allows our model to simulate and learn the diverse appearance changes of target instance in real-world scenarios during backward tracking, achieving temporal cross-frame association.

Algorithm 1: The SSTrack training process

Input: Initial frame and bounding box (I_r, B_r); Search frame $I_s^{t=2:n}$
Output: B_s^t in subsequent frames

// Forward tracking
for t = 2 to n **do**
  $B_s^t = \mathcal{E}\left(I_s^t, (I_r, B_r)\right)$
  Crop I_s^t based on B_s^t yields a new reference frame I_{sr}^t
end for
// Backward tracking
Expand I_r to get multiple new views $I_r^{1:m}$
for t = 1 to m **do**
  $B_r^t = \mathcal{E}\left(I_r^t, \{(I_{sr}, B_{sr})\}_n\right)$
end for
// Tracking and contrastive losses
Calculate loss using Eq.6 and update parameters.
return B_s^t

With this decoupled learning way, we achieve global spatial localization and local temporal association within a unified tracking framework. This effectively leverages both labeled and unlabeled video data, making it easy to learn target correspondences across timestamps from diverse scenarios. It is worth noting that the proposed decoupled spatio-temporal consistency training framework is used only during the training phase. To improve inference efficiency, we retain only the local (backward) tracking component for the inference model.

Instance Contrastive Learning

Another performance bottleneck of traditional self-supervised tracking algorithms is the difficulty in learning a robust instance representation. A mature visual tracker must accurately locate the target from diverse and complex backgrounds, requiring the ability to distinguish between different instances. However, directly applying contrastive learning methods (He et al. 2020; Oord, Li, and Vinyals 2018; Chen and He 2021) in self-supervised tracking to improve feature discriminability is not feasible, as they rely on clean, target-centered images. In other words, due to the highly open nature of instances and scenarios in the tracking domain, frame-level similarity learning is insufficient to distinguish instances in complex scenarios. Therefore, we introduce a simple yet effective instance contrastive loss function that mines rich instance information from a large number of unlabeled video sequences to efficiently learn diverse instance correspondences.

Given an initial frame $I_r \in \mathbb{R}^{3\times H_r \times W_r}$, we apply various data augmentation operations to generate different views, simulating the appearance changes of the object at different timestamps in the video, thereby automatically obtaining target correspondences. Since annotations are unknown in the contrastive learning process, we design an additional

Tracker	Type	Resolution	Params	FLOPs	Speed	Device
SeqTrack	ViT-B	384×384	89M	148G	$21 fps$	A100
AQATrack	HiViT-B	384×384	72M	58G	$57 fps$	A100
SSTrack	ViT-B	384×384	92M	73G	$59 fps$	A100

Table 1: Comparison of model parameters, FLOPs, and inference speed.

mask matrix $\mathcal{M}$ for each view based on the prediction results to extract the target instance from the background, where 1 represents the target region and 0 represents the background region. Subsequently, we perform a pooling operation to obtain the corresponding target representation, achieving instance supervision without any labels. Formally, our instance-level contrastive loss function is as follows:

$$\mathcal{L}_{cont} = -\sum_{q \in Q} \log \frac{\exp\big(\text{sim}(q, q^+)/\tau\big)}{\sum_{q^- \in Q^-} \exp(\text{sim}(q, q^-)/\tau)}, \quad (5)$$

where Q represents the set of all potential instances in a batch. q^+ and q^- denote the positive and negative samples to q, respectively. Positive samples are different views of the same instance obtained through various data augmentation operations. Negative samples come from different instances. Additionally, $\text{sim}(\cdot)$ denotes the cosine similarity between any sample pairs and τ is a temperature parameter.

Through this learning process, we make the representations of the same instance as similar as possible in the feature space while maximizing the distance between representations of different instances. This allows our model to effectively learn robust instance tracking representations in a self-supervised manner, reducing reliance on extensive box annotations. The overall optimization objective can be formulated as:

$$\mathcal{L}_{all} = \mathcal{L}_{track} + \mathcal{L}_{cont}. \quad (6)$$

Finally, we summarize the process of the proposed self-supervised tracking algorithm as shown in the Algorithm 1.

Experiments

Implementation Details

We use a ViT-Base (Dosovitskiy et al. 2021) model with DropMAE (Wu et al. 2023) pre-trained parameters as the visual encoder. The training data includes LaSOT (Fan et al. 2019), GOT-10k (Huang, Zhao, and Huang 2021), TrackingNet (Müller et al. 2018), and COCO (Lin et al. 2014). The AdamW (Loshchilov and Hutter 2019) is used to end-to-end optimize model parameters with initial learning rate of 2.5×10^{-5} for the backbone, 2.5×10^{-4} for the rest, and set the weight decay to 10^{-4}. The training epochs is set to 150 epochs. $10k$ image pairs are randomly sampled in each epoch. The learning rate drops by a factor of 10 after 120 epochs. The model is conducted on a server with two 80GB Tesla A100 GPUs and set the batch size to be 8. For forward tracking, we use one reference frame and three global search frames as the model input. For backward tracking, we use three reference frames and two cropped search frames as the input. The reference and search frames for backward tracking are derived from the search and reference frames of forward tracking, with the reference frames augmented from different views.

On the other hand, we analyze the parameters, FLOPs, and inference speed of different models. As shown in Tab.1, our SSTrack runs at 59 *fps* on an A100 GPU. Compared to SeqTrack (Chen et al. 2023) and AQATrack (Xie et al. 2024), we achieve faster inference speed.

Comparison with the SOTA

We compare the performance of our method with previous self-supervised tracking methods on the GOT10K (Huang, Zhao, and Huang 2021), LaSOT (Fan et al. 2019), TrackingNet (Müller et al. 2018), OTB100 (Wu, Lim, and Yang 2015), UAV123 (Mueller, Smith, and Ghanem 2016), and VOT2018 (Kristan et al. 2018) datasets. We then compare our tracker with more fully supervised methods on the $\text{LaSOT}_{\text{ext}}$ (Fan et al. 2021), TNL2K (Wang et al. 2021b), and VOT2020 (Kristan, Leonardis, and et.al 2020) datasets.

GOT10K. GOT10K is a popular general tracking benchmark containing over $10,000$ video sequences. Under the one-shot protocol of the GOT10K dataset, we compare our SSTrack with both self- and fully- supervised algorithms. As shown in the Tab.2, compared to the self-supervised method TADS, our tracker significantly outperforms by 25.7%, 27.1%, and 45.1% in AO, $\text{SR}_{0.5}$, and $\text{SR}_{0.75}$ metrics, respectively. Additionally, our method significantly narrows the performance gap with fully supervised methods. This performance gain is primarily attributed to the proposed decoupled spatio-temporal consistency training framework, which effectively leverages both labeled and unlabeled video data to learn the spatio-temporal context of target instance.

LaSOT. LaSOT is a classic long-term tracking benchmark, comprising 1120 training sequences and 280 test sequences. As shown in the Tab.2, compared to the self-supervised method TADS, our method improves the success, normalized precision, and precision score by 20.4%, 22.2%, and 25.9%, respectively. Additionally, compared to the state-of-the-art supervised method ODTrack, the AUC score gap of our self-supervised tracker is reduced to 7.3%. These results indicate that the proposed instance contrastive loss function helps the model learn the appearance and motion information of target, significantly enhancing self-supervised tracking performance in long-term scenarios.

TrackingNet. TrackingNet is a large-scale tracking benchmark with extensive manual bounding box annotations, offering a testset with 511 video sequences. As shown in Tab.2, our tracker achieves excellent results in short-term tracking scenarios. For example, compared to the self-supervised method TADS, SSTrack surpasses it by 14.8% in AUC score. Additionally, compared to the state-of-the-art supervised tracker ARTrackV2, SSTrack is only 3.5% lower in normalized precision, significantly enhancing the potential of self-supervised tracking algorithms.

OTB100, UAV123, and VOT2018. OTB100, UAV123, and VOT2018 are widely-used visual tracking datasets, comprising a variety of video sequences that pose challenges such as occlusion, lighting variations, motion changes, and camera motion. As indicated in Tab.3, compared to most self/fully-supervised tracking methods, our approach

Type	Method	GOT10K*			LaSOT			TrackingNet			$\text{LaSOT}_{\text{ext}}$		
		AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P
Fully Sup	SiamPRN++ (Li et al. 2019)	51.7	61.6	32.5	49.6	56.9	49.1	73.3	80.0	69.4	34.0	41.6	39.6
	DiMP (Bhat et al. 2019)	61.1	71.7	49.2	56.9	65.0	56.7	74.0	80.1	68.7	39.2	47.6	45.1
	SiamRCNN (Voigtlaender et al. 2020)	64.9	72.8	59.7	64.8	72.2	-	81.2	85.4	80.0	-	-	-
	Ocean (Zhang et al. 2020)	61.1	72.1	47.3	56.0	65.1	56.6	-	-	-	-	-	-
	STMTrack (Fu et al. 2021)	64.2	73.7	57.5	60.6	69.3	63.3	80.3	85.1	76.7	-	-	-
	TrDiMP (Wang et al. 2021a)	67.1	77.7	58.3	63.9	-	61.4	78.4	83.3	73.1	-	-	-
	TransT (Chen et al. 2021)	67.1	76.8	60.9	64.9	73.8	69.0	81.4	86.7	80.3	-	-	-
	Stark (Yan et al. 2021)	68.8	78.1	64.1	67.1	77.0	-	82.0	86.9	-	-	-	-
	KeepTrack (Mayer et al. 2021)	-	-	-	67.1	77.2	70.2	-	-	-	48.2	-	-
	SBT-B (Xie et al. 2022)	69.9	80.4	63.6	65.9	-	70.0	-	-	-	-	-	-
	Mixformer (Cui et al. 2022)	70.7	80.0	67.8	69.2	78.7	74.7	83.1	88.1	81.6	-	-	-
	TransInMo (Guo et al. 2022)	-	-	-	65.7	76.0	70.7	81.7	-	-	-	-	-
	OSTrack (Ye et al. 2022)	73.7	83.2	70.8	71.1	81.1	77.6	83.9	88.5	83.2	50.5	61.3	57.6
	AiATrack (Gao et al. 2022)	69.6	80.0	63.2	69.0	79.4	73.8	82.7	87.8	80.4	47.7	55.6	55.4
	SeqTrack (Chen et al. 2023)	74.5	84.3	71.4	71.5	81.1	77.8	83.9	88.8	83.6	50.5	61.6	57.5
	GRM (Gao, Zhou, and Zhang 2023)	73.4	82.9	70.4	69.9	79.3	75.8	84.0	88.7	83.3	-	-	-
	VideoTrack (Xie et al. 2023)	72.9	81.9	69.8	70.2	-	76.4	83.8	88.7	83.1	-	-	-
	ARTrack (Xing et al. 2023)	75.5	84.3	74.3	72.6	81.7	79.1	85.1	89.1	84.8	51.9	62.0	58.5
	EVPTrack (Shi et al. 2024)	76.6	86.7	73.9	72.7	82.9	80.3	84.4	89.1	-	53.7	65.5	61.9
	ODTrack (Zheng et al. 2024)	77.0	87.9	75.1	73.2	83.2	80.6	85.1	90.1	84.9	52.4	63.9	60.1
	HIPTrack (Cai, Liu, and Wang 2024)	77.4	88.0	74.5	72.7	82.9	79.5	84.5	89.1	83.8	-	-	-
	AQATrack (Xie et al. 2024)	76.0	85.2	74.9	72.7	82.9	80.2	84.8	89.3	84.3	52.7	64.2	60.8
	ARTrackV2 (Bai et al. 2024)	77.5	86.0	75.5	73.0	82.0	79.6	85.7	89.8	85.5	52.9	63.4	59.1
Self Sup	TADS (Li et al. 2023)	46.7	56.5	21.1	45.5	54.2	44.8	65.6	73.4	60.6	-	-	-
	SSTrack-256	67.1	76.6	59.1	64.8	75.2	69.7	80.1	**86.7**	**78.9**	46.2	57.8	52.1
	SSTrack-384	**72.4**	**83.6**	**66.2**	**65.9**	**76.4**	**70.7**	**80.4**	86.3	77.9	**48.5**	**60.9**	**54.5**

Table 2: Comparison with state-of-the-arts on four popular benchmarks: GOT10K, LaSOT, TrackingNet, and $\text{LaSOT}_{\text{ext}}$. Where $*$ denotes for trackers only trained on GOT10K. Best in **bold**, second best underlined.

Datesets	Fully Supervised Tracking							Self Supervised Tracking					
	ATOM	Ocean	DiMP	TransT	TrDiMP	Mixformer	HIPTrack	S^2SiamFC	CycleSiam	self-SDCT	TADS	**SSTrack-256**	**SSTrack-384**
OTB100(AUC)	67.1	68.4	68.4	69.4	67.5	70.0	71.0	-	-	63.8	65.3	67.9	**70.5**
UAV123(AUC)	64.3	-	65.3	69.1	67.5	70.4	70.5	-	-	50.1	55.2	65.5	**66.1**
VOT2018(Acc)	0.590	0.592	0.597	-	-	-	-	0.463	0.562	-	-	0.587	**0.630**

Table 3: Comparison with SOTA methods on OTB100, UAV123, and VOT2018 datasets. Best in **bold**, second best underlined.

achieves excellent results across the OTB dataset. Specifically, our SSTrack outperforms TADS by 10.9% in AUC score on the UAV123 dataset. Additionally, our tracker also surpasses CycleSiam by 6.8% in accuracy on the VOT2018 dataset. These results demonstrate that our tracker maintains excellent generalization across various tracking scenarios.

$\text{LaSOT}_{\text{ext}}$, TNL2K, and VOT2020. $\text{LaSOT}_{\text{ext}}$, TNL2K, and VOT2020 are large-scale tracking datasets that include more challenging video sequences. Most state-of-the-art supervised trackers are evaluated on these benchmarks to verify their accuracy and robustness. As shown in Tab.2, 4, and 5, our self-supervised tracking framework achieves competitive results and significantly narrows the performance gap with fully supervised methods. These results demonstrate the effectiveness of our proposed method, achieving good tracking performance even with limited annotations.

Ablation Study

Importance of decoupled spatio-temporal consistency training framework. As shown in Tab.6, *baseline* represents a self-supervised model based on contrastive learning. When we introduce our decoupled training framework, its performance significantly improves, achieving an increase of 26.8% in AO score. By incorporating spatio-temporal context into our self-supervised framework, the tracking performance improves by an additional 2.2% in AO score. These results indicate that the decoupled training framework effectively learns instance correspondences across various tracking scenarios, playing a crucial role in our self-supervised tracking framework.

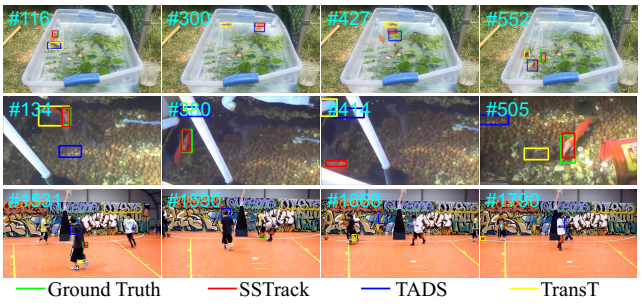

Figure 3: Qualitative comparison of our tracker with self- and fully- supervised trackers on LaSOT benchmark.

Importance of instance contrastive loss. As shown in

Metrics	Fully Supervised Tracking											Self Sup	
	SiamFC	MDNet	SiamRPN++	Ocean	TransT	OSTrack	SeqTrack	ARTrack	F-BDMTrack	ODTrack	AQATrack	**SSTrack-256**	**SSTrack-384**
AUC(%)	29.5	38.0	41.3	38.4	50.7	55.9	56.4	59.8	57.8	60.9	59.3	52.1	**53.8**
P(%)	28.6	37.1	41.2	37.7	51.7	-	-	-	-	64.5	62.3	53.3	**55.3**

Table 4: Comparison with state-of-the-art methods on TNL2K benchmark. Our results are in **bold** and underline.

Metrics	Fully Supervised Tracking											Self Sup	
	STM	SiamMask	Ocean	D3S	AlphaRef	Ocean+	STARK	SBT	Mixformer	SeqTrack	ODTrack	**SSTrack-256**	**SSTrack-384**
EAO(↑)	0.308	0.321	0.430	0.439	0.482	0.491	0.505	0.515	0.535	0.522	0.581	0.458	**0.503**
Accuracy(↑)	0.751	0.624	0.693	0.699	0.754	0.685	0.759	0.752	0.761	-	0.764	0.664	**0.754**
Robustness(↑)	0.574	0.648	0.754	0.769	0.777	0.842	0.819	0.825	0.854	-	0.877	0.839	**0.816**

Table 5: Comparison with state-of-the-art methods on VOT2020 benchmark. Our results are in **bold** and underline.

#	Method	AO	$SR_{0.5}$	$SR_{0.75}$
1	*Baseline*	41.8	45.2	14.3
2	*Decoupled training w/o context*	68.6	78.1	60.1
3	*+ Spatio-temporal context*	70.8	82.2	64.6
4	*+ Instance contrastive loss*	72.4	83.6	66.2

Table 6: Ablation studies of our tracker variants in GOT-10k.

#	Pre-trained Model	AO	$SR_{0.5}$	$SR_{0.75}$
1	*MAE*	62.7	71.1	54.1
2	*DropMAE*	68.6	78.1	60.1

Table 7: Comparison of different pre-trained models.

the fourth row of Tab.6, by adding the instance contrastive loss to our self-supervised framework, our tracker improves by 1.6%, 1.4%, and 1.6% in AO, $SR_{0.5}$, and $SR_{0.75}$ metrics, respectively. This result demonstrates that the instance contrastive loss effectively enhances the discriminability of instance representations in self-supervised learning, thereby improving the tracking performance of our model.

Effect of pre-trained model. We conduct experiments in Tab.7 to validate the impact of different pre-trained models on our model. We observe that by replacing the MAE pre-trained model with DropMAE, our model achieves a performance gain of 5.9% in AO score. This demonstrates that DropMAE, with temporal correspondence learning, is more suitable for self-supervised tracking task and more effectively learns target correspondences from unlabeled data.

Effect of data augmentations. Due to the absence of annotations, self-supervised algorithms struggle to achieve sufficient training. To address this issue, we incorporate various data augmentation methods such as shear, blur, and LSJ (Ghiasi et al. 2021). As shown in Tab.8, these augmentation methods enhance our model's ability to learn target variations from different views by increasing data diversity. In other words, appropriate data augmentation methods improve our model's robustness.

Visualization and Limitation

Visualization. Furthermore, we conduct qualitative experiments to visually demonstrate the effectiveness of the proposed framework. Fig.3 presents a visual comparison of our SSTrack with the advanced self-supervised tracker TADS and the fully-supervised tracker TransT. By effectively learning the spatio-temporal context of object in a self-supervised manner and improving the instance features discriminability through contrastive learning, our model performs exceptionally well in various complex scenarios, even rivaling fully supervised tracking algorithms.

#	Shear	Blur	LSJ	AO	$SR_{0.5}$	$SR_{0.75}$
1	✓	✓	✓	70.3	81.2	63.3
2	✗			69.4	80.5	62.4
3		✗		69.3	79.8	61.3
4			✗	69.0	80.0	61.6
5				68.6	78.1	60.1

Table 8: Comparison of different data augmentations.

Limitation. This work leverages the powerful advantages of the decoupled spatio-temporal consistency training framework and instance contrastive learning to design a novel self-supervised tracking algorithm. Despite achieving remarkable results, we observe that the performance of backward tracking somewhat depends on the localization accuracy of forward tracking. Thus, improving the accuracy of forward tracking could further enhance the performance of self-supervised tracking, potentially narrowing the gap with fully supervised tracking algorithms even more.

Conclusion

In this work, we have proposed a self-supervised tracking framework named SSTrack, aimed at eliminating the reliance on costly box annotations. Specifically, we have introduced a simple yet efficient decoupled spatio-temporal consistency training framework to learn rich target appearance and motion information across timestamps. Furthermore, we have proposed an instance contrastive loss function to learn instance-level correspondences from a multi-view perspective, providing reliable instance supervision without any labels. Extensive experiments have demonstrated the superiority of our method. We hope this work will further inspire research into self-supervised tracking algorithms.

Acknowledgements

This work is supported by the Project of Guangxi Science and Technology (No.2024GXNSFGA010001 and 2022GXNSFDA035079), the National Natural Science Foundation of China (No.U23A20383 and 62472109), the Guangxi "Young Bagui Scholar"Teams for Innovation and Research Project, the Research Project of Guangxi Normal University (No.2024DF001), and the Innovation Project of Guangxi Graduate Education (YCBZ2024083).

References

Bai, Y.; Zhao, Z.; Gong, Y.; and Wei, X. 2024. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19048–19057.

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. S. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In *ECCV Workshops*, 850–865.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning Discriminative Model Prediction for Tracking. In *ICCV*, 6181–6190.

Cai, W.; Liu, Q.; and Wang, Y. 2024. HIPTrack: Visual Tracking with Historical Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19258–19267.

Chen, B.; Li, P.; Bai, L.; Qiao, L.; Shen, Q.; Li, B.; Gan, W.; Wu, W.; and Ouyang, W. 2022. Backbone is All Your Need: A Simplified Architecture for Visual Object Tracking. In *ECCV (22)*, 375–392.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. *CVPR*, abs/2304.14394.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer Tracking. In *CVPR*, 8126–8135.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese Box Adaptive Network for Visual Tracking. In *CVPR*, 6667–6676.

Cheng, S.; Zhong, B.; Li, G.; Liu, X.; Tang, Z.; Li, X.; and Wang, J. 2021. Learning To Filter: Siamese Relation Network for Robust Tracking. In *CVPR*, 4421–4431.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2022. MixFormer: End-to-End Tracking with Iterative Mixed Attention. In *CVPR*, 13598–13608.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fan, H.; Bai, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Harshit; Huang, M.; Liu, J.; Xu, Y.; Liao, C.; Yuan, L.; and Ling, H. 2021. LaSOT: A High-quality Large-scale Single Object Tracking Benchmark. *Int. J. Comput. Vis.*, 439–461.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In *CVPR*, 5374–5383.

Fu, Z.; Liu, Q.; Fu, Z.; and Wang, Y. 2021. STMTrack: Template-Free Visual Tracking With Space-Time Memory Networks. In *CVPR*, 13774–13783.

Gao, S.; Zhou, C.; Ma, C.; Wang, X.; and Yuan, J. 2022. AiATrack: Attention in Attention for Transformer Visual Tracking. In *ECCV (22)*, 146–164.

Gao, S.; Zhou, C.; and Zhang, J. 2023. Generalized Relation Modeling for Transformer Tracking. *CVPR*, abs/2303.16580.

Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.-Y.; Cubuk, E. D.; Le, Q. V.; and Zoph, B. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2918–2928.

Guo, M.; Zhang, Z.; Fan, H.; Jing, L.; Lyu, Y.; Li, B.; and Hu, W. 2022. Learning Target-aware Representation for Visual Tracking via Informative Interactions. In *IJCAI*, 927–934.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Huang, L.; Zhao, X.; and Huang, K. 2021. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5): 1562–1577.

Kristan, M.; Leonardis, A.; and et.al. 2020. The Eighth Visual Object Tracking VOT2020 Challenge Results. In *ECCV Workshops (5)*, volume 12539 of *Lecture Notes in Computer Science*, 547–601. Springer.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R.; ˇCehovin Zajc, L.; Vojir, T.; Bhat, G.; Lukezic, A.; Eldesokey, A.; et al. 2018. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *CVPR*, 4282–4291.

Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018. High Performance Visual Tracking With Siamese Region Proposal Network. In *CVPR*, 8971–8980.

Li, X.; Pei, W.; Wang, Y.; He, Z.; Lu, H.; and Yang, M.-H. 2023. Self-supervised tracking via target-aware data synthesis. *IEEE Transactions on Neural Networks and Learning Systems*.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *ICCV*, 2999–3007.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, 740–755.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Mayer, C.; Danelljan, M.; Paudel, D. P.; and Gool, L. V. 2021. Learning Target Candidate Association to Keep Track of What Not to Track. In *ICCV*, 13424–13434. IEEE.

Mueller, M.; Smith, N.; and Ghanem, B. 2016. A Benchmark and Simulator for UAV Tracking. In *ECCV*, 445–461.

Müller, M.; Bibi, A.; Giancola, S.; Al-Subaihi, S.; and Ghanem, B. 2018. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *ECCV*, 310–327.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*, 658–666.

Shi, L.; Zhong, B.; Liang, Q.; Li, N.; Zhang, S.; and Li, X. 2024. Explicit Visual Prompts for Visual Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4838–4846.

Sio, C. H.; Ma, Y.-J.; Shuai, H.-H.; Chen, J.-C.; and Cheng, W.-H. 2020. S2siamfc: Self-supervised fully convolutional siamese network for visual tracking. In *Proceedings of the 28th ACM international conference on multimedia*, 1948–1957.

Voigtlaender, P.; Luiten, J.; Torr, P. H. S.; and Leibe, B. 2020. Siam R-CNN: Visual Tracking by Re-Detection. In *CVPR*, 6577–6587.

Wang, N.; Zhou, W.; Wang, J.; and Li, H. 2021a. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, 1571–1580.

Wang, X.; Jabri, A.; and Efros, A. A. 2019. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2566–2576.

Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; and Wu, F. 2021b. Towards More Flexible and Accurate Object Tracking With Natural Language: Algorithms and Benchmark. In *CVPR*, 13763–13773.

Wu, Q.; Yang, T.; Liu, Z.; Wu, B.; Shan, Y.; and Chan, A. B. 2023. DropMAE: Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks. *CVPR*, abs/2304.00571.

Wu, Y.; Lim, J.; and Yang, M. 2015. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9): 1834–1848.

Xie, F.; Chu, L.; Li, J.; Lu, Y.; and Ma, C. 2023. VideoTrack: Learning to Track Objects via Video Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22826–22835.

Xie, F.; Wang, C.; Wang, G.; Cao, Y.; Yang, W.; and Zeng, W. 2022. Correlation-Aware Deep Tracking. In *CVPR*, 8741–8750.

Xie, J.; Zhong, B.; Mo, Z.; Zhang, S.; Shi, L.; Song, S.; and Ji, R. 2024. Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19300–19309.

Xing, W.; Yifan, B.; Yongchao, Z.; Dahu, S.; and Yihong, G. 2023. Autoregressive Visual Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9697–9706.

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021. Learning Spatio-Temporal Transformer for Visual Tracking. In *ICCV*, 10428–10437.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. In *ECCV (22)*, 341–357.

Yuan, D.; Chang, X.; Huang, P.-Y.; Liu, Q.; and He, Z. 2020. Self-supervised deep correlation tracking. *IEEE Transactions on Image Processing*, 30: 976–985.

Yuan, W.; Wang, M. Y.; and Chen, Q. 2020. Self-supervised object tracking with cycle-consistent siamese networks. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10351–10358. IEEE.

Zhang, Z.; and Peng, H. 2019. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In *CVPR*, 4591–4600.

Zhang, Z.; Peng, H.; Fu, J.; Li, B.; and Hu, W. 2020. Ocean: Object-Aware Anchor-Free Tracking. In *ECCV*, 771–787.

Zheng, Y.; Zhong, B.; Liang, Q.; Mo, Z.; Zhang, S.; and Li, X. 2024. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7588–7596.